03 020
YOR

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００３年　２月２７日

出願番号
Application Number: 特願２００３－０５０９１６

[ST.10/C]: [ＪＰ２００３－０５０９１６]

出願人
Applicant(s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２００３年　５月２０日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０３７０３０

【書類名】　特許願
【整理番号】　JP9030020
【提出日】　平成15年 2月27日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 17/00
G06F 17/22
【発明者】
【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
【氏名】　渋谷　哲朗
【特許出願人】
【識別番号】　390009531
【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション
【代理人】
【識別番号】　100086243
【弁理士】
【氏名又は名称】　坂口　博
【代理人】
【識別番号】　100091568
【弁理士】
【氏名又は名称】　市位　嘉宏
【代理人】
【識別番号】　100108501
【弁理士】
【氏名又は名称】　上野　剛史
【復代理人】
【識別番号】　100110607
【弁理士】

【氏名又は名称】　間山　進也

【手数料の表示】

【予納台帳番号】　062651

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0207860

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　塩基配列のスクリーニングを行うためのコンピュータ・システム、そのための方法、該方法をコンピュータに対して実行させるためのプログラムおよび該プログラムを記憶したコンピュータ可読な記録媒体

【特許請求の範囲】

【請求項１】　塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システム。

【請求項２】　前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含む、

請求項１に記載のコンピュータ・システム。

【請求項３】　前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断する、

請求項２に記載のコンピュータ・システム。

【請求項４】　塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システム。

【請求項５】　前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含む、

請求項４に記載のコンピュータ・システム。

【請求項６】　前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断する、

請求項５に記載のコンピュータ・システム。

【請求項７】　前記評価処理部は、前記動的停止判断部の判断に応答して、前記部分配列指定部に対して直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定させる、

請求項５に記載のコンピュータ・システム。

【請求項８】　コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法。

【請求項９】　前記評価ステップは、さらに前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断結果に応答して停止信号を生成するステップと

を含む請求項８に記載のコンピュータ制御方法。

【請求項１０】　コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法。

【請求項１１】　前記評価ステップは、少なくとも前記部分配列のうちの１つについて前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断の結果に応答して動的に評価を停止するステップとを含む、

請求項１０に記載のコンピュータ制御方法。

【請求項１２】　前記評価ステップは、さらに評価の停止を判断するステップと、

前記評価の停止の判断に応答して、直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定するステップと

を含む、請求項１１に記載のコンピュータ制御方法。

【請求項１３】　コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラム。

【請求項１４】　前記評価ステップは、さらに前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断結果に応答して評価を停止させるステップと

を含む請求項１３に記載のコンピュータ実行可能なプログラム。

【請求項１５】　コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラム。

【請求項１６】　前記評価ステップは、少なくとも前記部分配列のうちの１つについて前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断の結果に応答して動的に評価を停止するステップと、

前記評価の停止の判断に応答して、直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定するステップと

を含む、請求項１５に記載のコンピュータ実行可能なプログラム。

【請求項１７】　請求項１３に記載のプログラムを記録したコンピュータ可読な記憶媒体。

【請求項１８】　請求項１５に記載のプログラムを記録したコンピュータ可読な記憶媒体。

【請求項１９】　塩基配列のスクリーニングをネットワークを介して実行するためのサーバであって、該サーバは、

ターゲット塩基配列データを格納するデータベースと、

ネットワークを介してターゲット指定データとターゲット塩基配列に結合する可能性のあるプローブ塩基配列を受け取ると共に、スクリーニングの結果を前記ネットワークを介して送出する送受信部と、

前記プローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット指定データを使用して前記データベースを検索し、検索されたターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行すると共に、前記送受信部に伝送する評価処理部と、

を含むサーバ。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、塩基配列のスクリーニングに関し、より詳細には所定の塩基配列が、他の塩基配列の一部分を構成するか否かを効率よく判定するために、所定の塩基配列をスクリーニングするためのシステム、そのための方法、該方法をコンピュータに対して実行させるためのプログラム、該プログラムを記憶したコンピュータ可読な記録媒体およびネットワークを介して塩基配列のスクリーニングを実行させるためのサーバに関する。

【０００２】

【従来の技術】

ＤＮＡは、Ａ（アデニン）、Ｔ（チミン）、Ｃ（シトシン）、Ｇ（グアニン）からなる４種類の塩基から構成される塩基鎖であり、ＲＮＡはＡ、Ｕ（ウラシル）、Ｇ、Ｃからなる４種類の塩基から構成されることが知られている。これらの合計５種類の塩基は、相補塩基と呼ばれる組み合わせでそれぞれ特異的に結合することが知られており、アデニン（Ａ）とチミン（Ｔ）（ＲＮＡの場合にはアデニン（Ａ）とウラシル（Ｕ））、シトシン（Ｃ）とグアニン（Ｇ）とが、それぞれ結合し、２重鎖を形成することが知られている。この２重鎖がハイブリダイゼーションとして参照されており、２重鎖を生成することをハイブリダイズする、として参照されている。また、所与の塩基配列に対して、２重鎖を形成するように結合することが可能な塩基配列は、相補配列として参照されている。

【０００３】

上述した各塩基間の相補性という性質を利用することにより、所定のＤＮＡや、ＲＮＡを含む可能性のある試料中において、目的とするＤＮＡやＲＮＡ（以下、本発明においては、被検査対象とされる塩基配列のことを、ターゲット塩基配列として参照する。）が存在するか否か、を判断することが必要とされる場合がある。さらに、本発明においてはプローブとして参照され、ＤＮＡやＲＮＡの所与の領域の塩基に対して特異的に結合する塩基鎖を用意し、この塩基鎖がターゲット配列に対して結合、すなわちハイブリダイゼーションするかどうかを判断する必要が生じることもある。上述した判断を、結合性判断またはハイブリダイゼーション判断として参照する。また、ターゲット塩基配列を検出するために使用する相補配列をプローブとして参照する。

【０００４】

上述した結合性の判断は、種々の目的のために利用されており、例えばＤＮＡチップは、上述したハイブリダイゼーションの性質を利用する情報処理チップである。情報処理チップは、多くの場合様々なＤＮＡ、ＲＮＡといった塩基配列に対する相補配列を用意し、同時に大量のハイブリダイゼーション判断を行わせることにより、目的とする処理を実行する。また、いわゆるＰＣＲ（Polymerase Chain Reaction）は、ＤＮＡの配列上の２箇所の部分の配列に対して相補配列を作成し、これらの相補配列で囲まれる領域を大量にコピーすることにより、ＤＮＡの配列に対する判断・評価を行う手法である。

【０００５】

これらの相補配列を使用する判断・評価において多くの場合、対象とされる現実の試料に目的とは異なる別のＤＮＡやＲＮＡの配列が混在する場合がある。このような場合、作成または提供されるプローブが、混在する塩基配列には結合しないことを保証することにより、判断・評価の効率化・高精度化、および高信頼化を達成することが可能となる。また、所定のＤＮＡに対して特異的なプローブＰを作成するためにＤＮＡやＲＮＡのシンセサイザが使用される場合もあり、目的以外のプローブＰを効率よく排除してゆくことにより、タンパク合成およびスクリーニングを含めた効率は、著しく高められるものと考えられる。

【０００６】

図１８は、ターゲット塩基配列と、プローブとの関係を示した図である。図１８においては、ターゲット塩基配列が符号Ｔで示されており、プローブが符号Ｐで示されている。ターゲット塩基配列Ｔは、例えば数千の塩基が結合した長鎖の塩基配列とすることができる（以下、本発明においては塩基数のことを、base pair: bpとして記述する。）。本質的には図１８に示されるプローブ１０２は、ターゲット塩基配列Ｔのうち、Ｔｐで示される領域の配列の完全な相補配列となっていることが理想的である。

【０００７】

しかしながら、所定の塩基配列と別の塩基配列とは、現実的には完全な相補配列でなくても結合する性質がある（８０％－９０％の一致でも結合する）。また

、シーケンサなどの塩基配列決定装置における分析誤差もあるので、必ずしも１００％完全に一致しないからといって、プローブの候補となり得ないとして排除することは妥当ではない。このため、プローブＰとターゲット塩基配列とが結合しないことを確認するためには、従来は、Smith-Waterman法などの高精度アラインメント・アルゴリズムを適用して相互の塩基配列の解析を行い、プローブの相補配列に近い配列がターゲット塩基配列に全くないことを保証する必要とされていた。

【０００８】

また、これまで相互に近似する塩基配列を検索する手法として、BLAST (Altchul SF., Miller, GW., Myers EW., Lipman DJ., “Basic local alignment search tool”, J. Mol. Biol. 1990, Oct. 5, 215 (3), 403-410)、FASTA (Pearson, WR., Lipman, DJ., “Improved tools for biological sequence comparison”, Proc. Natl. Acad. Sci. USA, 1988, Apr; 85(8), 2444-2448 Related Articles, Links)などの高速検索アルゴリズムが提案されている。しかしながら、これらのアルゴリズムでも上述したようなターゲット塩基配列中における類似する部分配列を全部発見できるわけではなく、ターゲット配列に対してプローブが結合しないという保証を行うスクリーニング用途には適するものではない。

【０００９】

【発明が解決しようとする課題】

本発明は、従来技術の上述した不都合に鑑み、プローブＰが与えられた場合に、プローブＰとターゲット塩基配列Ｔとが少なくとも結合しないという評価の結果を、ユーザに対して迅速に提供することを可能とするものである。すなわち、プローブＰとターゲット塩基配列Ｔとが少なくとも結合しないことをユーザに対して示すことができれば、迅速にプローブの有効性および信頼性についての情報をユーザに対して提供することが可能となる。これまで種々の検索手法が提案されていたものの、ユーザの許容する範囲の誤差で所与のターゲット塩基配列に対して結合しないこと、または結合するという情報に加え、結合する場合には、可能性のあるプローブＰの配列をユーザに対して提示することを可能とする塩基配列のスクリーニング・システム、そのための方法、該方法をコンピュータに対し

て実行させるためのプログラム、該プログラムを記録したコンピュータ可読な記憶媒体、およびそのためのサーバが必要とされていた。

【００１０】

【課題を解決するための手段】

本発明は、ユーザが要求する精度内ではプローブＰとターゲット塩基配列Ｔとが少なくとも結合しないことをおよびユーザが指定する許容範囲内で互いに結合する可能性がある場合には、結合する可能性と、相似度および該当するサブシーケンスの情報をユーザに対して提示することができれば、高速なプローブのスクリーニングが可能になるという着想のもとになされたものである。

【００１１】

本発明の上記課題を解決するために、本発明においては、ユーザが最大の編集距離を指定し、最大編集距離だけ相違するか否かの評価から評価を始め、降順にプローブＰとターゲット塩基配列Ｔとの結合性についての評価を行う。その結果、ユーザが指定する最大編集距離よりも編集距離が大きくなることが明らかになった時点でターゲット塩基配列全体の評価を行わずに「結合しない」ことについての情報を提供することができる。さらに、本発明においては、評価停止条件をサブシーケンスと相補配列Ｑとの間の塩基差に対応させて動的に変更させることにより、結合する可能性があると判断された場合には、それらの相似度、塩基配列といった付帯的データをユーザに対して提供することが可能となる。

【００１２】

さらに、本発明においてはさらに精度を向上させ、より長鎖のターゲット塩基配列に対しても高速化、高精度化を確実に提供することを可能とする。すなわち、本発明の他の態様では、ターゲット塩基配列は、一定の重複を持つ部分配列Ｔｐに分割され、部分配列Ｔｐごとにプローブと結合しないことを判定する。より具体的には、評価したいターゲット塩基配列Ｔを適当なオーバーラップを与えた複数の部分配列Ｔｐ（ｐ＝１～ｗ）に分割する。分割された部分配列Ｔｐごとに、プローブＰの配列の相補配列が部分配列Ｔｐに含まれるか否かの判断を実行することで、さらに信頼性を高めることが可能となる。また、ターゲット塩基配列を部分配列Ｔｐに分割する場合の鎖長について、本発明の方法を使用すること

ができる。

【００１３】

すなわち、本発明によれば、塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システムが提供される。

【００１４】

本発明の前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含むことが好ましい。

【００１５】

本発明の前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断することができる。

【００１６】

本発明によれば、塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プロ

ーブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システムが提供される。

【００１７】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法が提供される。

【００１８】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法が提供される。

【００１９】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラムが提供される。

【００２０】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラムが提供される。

【００２１】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるコンピュータ制御方法のためのコンピュータ実行可能なプログラムが記録されたコンピュータ可読な記憶媒体であって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順から結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ可読な記憶媒体が提供される。

【００２２】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるコンピュータ制御方法のためのコンピュータ実行可能なプログラムが記録されたコンピュータ可読な記憶媒体であって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集

距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ可読な記憶媒体が提供される。

【００２３】

本発明によれば、塩基配列のスクリーニングをネットワークを介して実行するためのサーバであって、該サーバは、

ターゲット塩基配列データを格納するデータベースと、

ネットワークを介してターゲット指定データとターゲット塩基配列に結合する可能性のあるプローブ塩基配列を受け取ると共に、スクリーニングの結果を前記ネットワークを介して送出する送受信部と、

前記プローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット指定データを使用して前記データベースを検索し、検索されたターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行すると共に、前記送受信部に伝送する評価処理部と、

を含むサーバが提供される。

【００２４】

【発明の実施の形態】

以下、本発明につき図面に示した具体的な実施の形態をもって説明するが、本発明は、後述する実施の形態に限定されるものではない。本発明における結合性判断において、ターゲット塩基配列から得られるコンピュータ処理可能なターゲット塩基配列データを、以下、ターゲット塩基配列Ｔとして参照し、プローブである塩基配列から得られるコンピュータ処理可能な塩基配列データを、プローブＰとして参照する。さらに、プローブＰのコンピュータ処理可能な相補配列デー

タを相補配列Ｑとして参照する。

【００２５】

Ａ：本発明における使用パラメータおよび本質的処理の概説

図１は、ターゲット塩基配列Ｔと、プローブＰと、相補配列Ｑとの関連性を示した図である。図１（ａ）に示すように、本発明の特定の実施の形態において、プローブＰをＰ＝ＡＴＧＧＣＣＴＧＡＣとすると、Ｐの相補配列Ｑは、Ｑ＝ＧＴＣＡＧＧＣＣＡＴとなる。図１（ａ）に示すようにプローブＰと相補配列Ｑとの関係は、各塩基を相補塩基に置き換えただけではなく、配列の向きが図１（ａ）に示される矢線ALにしたがい、逆となっていることに留意されたい。また、相補配列Ｑは、ユーザが指定する最大編集距離内においてターゲット塩基配列Ｔの一部を構成する配列部分となっていることが、本発明の評価の本質的な処理となる。

【００２６】

このとき、本発明においてはターゲット塩基配列Ｔと、プローブＰまたは相補配列Ｑとの間に相似度ｒを定義する。本発明においては、ターゲット塩基配列Ｔとプローブとが結合するために、少なくともターゲット塩基配列Ｔが、プローブＰとの相似度でｒ以上の配列部分を含むことを要求する。すなわち、本発明では、ターゲット塩基配列Ｔとプローブとは、ターゲット塩基配列中に相似度ｒの配列部分がなければ、結合できないものとされる。

【００２７】

相似度ｒを決定するためには種々の方法を使用することができ、本発明において具体的には、以下の基準を使用して相似度ｒを決定することができる。ターゲット塩基配列Ｔに含まれるサブシーケンスＳとすると、サブシーケンスＳから相補配列Ｑを生成するために、塩基の置換、欠失、挿入といった文字解析・判断処理が合計で最低でｋ回必要であるものとする。値ｋは、サブシーケンスＳと相補配列Ｑとの間の最大編集距離として参照することができる。ここで、プローブＰに対する相補配列Ｑの塩基鎖長をmとして定義すれば、相似度rは、r=(m-k)/mとして定義することができる。したがって相似度がr以上とは、rについての上記式を参照して、最大編集距離がm(1-r)以下のサブシーケンスＳと相補配列Ｑとを指

定することに対応する。

【００２８】

図１（ｂ）は、上述した編集距離の決定する処理を具体的な実施の形態を使用して説明した図である。図１（ｂ）には、サブシーケンスＳ＝ＡＴＧＣＧＣＧＴＡＡＧＴであって、相補配列Ｑ＝ＡＴＧＣＧＴＴＡＡＣＴである場合の実施の形態をもって説明されている。図１（ｂ）に示した実施の形態では、サブシーケンスＳと、相補配列Ｑとの間の編集距離は4となり、相似度ｒは、r=(11-4)/11=0.6364となる。図１（ｂ）には、部分配列Ｔｐと相補配列Ｑとをそれぞれ整列（以下、上述した処理を、アラインメントとして参照する。）させて示している。図１（ｂ）に示した実施の形態では、欠失が２回あり、置換および挿入がそれぞれ１回なので、編集距離は４である。

【００２９】

本発明においては、所定のターゲット塩基配列Ｔにおいて、ターゲット塩基配列Ｔと、相補配列Ｑとの間には、ターゲット塩基配列Ｔに相補配列Ｑの(m-k)個以上の塩基が順番を違えずに出現することがない場合、ターゲット塩基配列Ｔには相補配列Ｑとの編集距離がk以下のものは存在しないということができる。プローブＰと相補配列Ｑとの間には、図１（ａ）に示した関係が存在するため、プローブＰから生成された相補配列Ｑとターゲット塩基配列Ｔについての情報を得ることができれば、プローブＰとサブシーケンスＳとについての関係を容易に得ることが可能となる。

【００３０】

一方で、本発明においては、所与のプローブＰが結合してしまう可能性のあるターゲット塩基配列をスクリーニングすることもできる。図２は、本発明において部分配列Ｔｐと相補配列Ｑとの関係において、相補配列Ｑの特異性を判断する場合の処理を示した図である。図２では、複数のターゲット塩基配列がスクリーニングの対象とされており、プローブＰがターゲット塩基配列T_1のみに特異的に結合するか否かを判断する実施の形態の概略図である。ターゲット塩基配列T_1には、プローブＰの相補配列Ｑがそのまま含まれているのが示されており、またT_aについても、相補配列Q_aが、ちょうど(m-a)の編集距離（ａは、ｋよりも

大きい正の正数である。）で存在しているのが示されている。また、ターゲット配列T_bには、(m-b)（ｂは、ｋよりも小さい正の整数である。）の編集距離部分配列が存在しており、この結果、プローブは、ターゲット塩基配列T_1およびT_bについて結合する可能性があることが示される。この場合、本発明によれば、結合可能性の情報と共にサブシーケンスとの相似度および塩基鎖長がユーザに提示されるので、プローブＰをさらに特性を付与するべく設計することも可能である。

【００３１】

Ｂ．本発明のスクリーニング方法の処理

上述した特性を効率的に利用することにより、本発明では、ターゲット塩基配列が、プローブＰとして適切か否か、また他のプローブＰを提供する必要があるか否かの判断を、高精度な逐次的解析するよりも、より高速に確認することが可能となる。すなわち、本発明によれば、ターゲット塩基配列Ｔにプローブ Ｐがハイブリダイズしないことの保証が可能になる。図３には、本発明の方法の本質的な処理のフローチャートを示す。図３に示すように、本発明の塩基配列のスクリーニング方法は、ステップＳ１０においてプローブＰとするべき塩基配列およびターゲット塩基配列Ｔを受け取る。この際の入力としては、シーケンサなどによって実際に得られた塩基配列でも良い。さらに、プローブＰは、塩基配列のシンセサイザなどによって合成することもできる。また、ターゲット塩基配列については、例えばＤＮＡまたはＲＮＡといった塩基配列を格納したデータベースから、インタラクティブにユーザがターゲット塩基配列を選択することもできるし、ほ乳類、真核微生物などのlive tissueから採取したＤＮＡサンプルを使用することもできる。図３における本発明のスクリーニング方法は、次いで、ステップＳ１２へと進み、プローブＰから相補配列Ｑを生成し、適切な記憶領域に格納すると共に、最大編集距離ｋを受け取り、適切な記憶領域に格納する。

【００３２】

ステップＳ１４では、検査するべきターゲット塩基配列Ｔについて、関数largest_prefix_with_common_subsequence()を適用して、その時点で検討を加えているターゲット塩基配列Ｔについて、相補配列Ｑがターゲット塩基配列Ｔのサブシ

ーケンスとして存在するか否かを判断することにより、結合性の評価を行なう。さらの本発明の処理は、ステップＳ１６へと進み、評価結果をコンピュータにおける適切な記憶領域に格納する。さらにステップＳ１６では、格納された評価結果を読み出して「結合する」、「結合しない」の判断を実行し、その結果がユーザに返され、ユーザに対して判断の結果が提供される。

【００３３】

図４は、本発明のスクリーニング方法の第２の実施の形態の処理を示したフローチャートである。図４に示した実施の形態では、対象とされるターゲット塩基配列Ｔは、複数T_1～T_wとして存在し、ターゲット塩基配列ごとに評価結果が作成され、記憶領域に格納される。図４に示した実施の形態ではステップＳ２０において、プローブＰとするべき塩基配列および複数のターゲット塩基配列T_1～T_wを受け取り、それぞれの格納部に格納する。ステップＳ２２において、プローブＱから相補配列Ｑを決定し、さらに編集距離ｋ、評価を繰り返して実行する部分配列の長さの単位(length)と、重なり合いの長さ(overlap_length)といった変数の入力を受け取る。ステップＳ２４においてカウンタtを、t=１に設定し、ステップＳ２６においてターゲット塩基配列に対して結合性の評価を実行する。ステップＳ２６においてターゲット塩基配列について評価を行い、その結果を記憶部に格納する。ステップＳ２８では、評価結果にアクセスして結合可能なサブシーケンスの有無などの情報を生成する。

【００３４】

ステップＳ３０においてカウンタtをインクリメントし、ステップＳ３２において、tがwよりも小さいか否かを判断する。ステップＳ３２の判断においてt≦wの場合(yes)には、ステップＳ３４に進み、T_{t+1}のターゲット塩基配列を記憶領域から読み出し、ステップＳ２６へと戻り、T_{t+1}のターゲット塩基配列についての評価結果を作成して格納する。ステップＳ３２の判断がt＞w場合(no)には、すでに評価を行うターゲット塩基配列について評価を終了しているので、ステップＳ３６へと進み、評価結果を判断してユーザに結合性の判断結果を提示する。

【００３５】

図５は、本発明において結合性を評価する処理を実行する関数である、largest_prefix_with_common_subsequence()のフローチャートを示した図である。図５に示した本発明における結合性の評価の処理は、ステップＳ４０の初期化処理から開始し、positions[i]の値および編集距離カウンタiの値を初期化する。ここで、positions[i]は、関数next_char_position()で与えられる値を返すか、または関数next_char_position()で与えられる値とpositions[i-1]で与えられる値のうちの小さい方の値を返す関数である。また、関数next_char_position()は、引数としてターゲット塩基配列、編集距離カウンタi、相補配列に存在する塩基の文字cとを引数として使用し、next_char_position(T, i, c)は、ターゲット塩基配列Ｔの塩基鎖の先頭から数えて、T[i..n]の中で最初に出てくる文字cの位置を返す関数である。また、文字cの塩基がT[i..n]の中に出現しない場合には、値(n+1)を返す関数である。

【００３６】

ステップＳ４２においては、評価を行う相補塩基の塩基位置を示す変数であるprobe_positionの値およびmin_kの値を初期化する。このmin_kの値は、min_kよりも小さな編集距離はありえない値を規定するものであり、部分配列と相補配列Ｑとの間の塩基の不致の最小値を与える尺度である。すなわち、min_k=0は、サブシーケンスＳと、相補配列Ｑとが完全に一致することに対応する。同時にmin_kは、停止判断を行うための停止条件値としても使用され、さらにプローブＰとサブシーケンスＳとの間の塩基誤差の尺度を与える。ステップＳ４４においては、相補塩基Ｑに含まれる塩基数に対応する変数probe_positionが相補塩基Ｑの塩基数であるmよりも大きいか否かを判断する。ステップＳ４４における判断において変数probe_positionがmよりも大きくなった場合(yes)には、すでに評価するべき相補塩基の塩基は残されていないので、ステップＳ４８へと進む。この段階では、編集距離カウンタiはmin_kよりも大きいので、m-k以上の鎖長のサブシーケンスが存在する可能性がまだ残されているので、positions[k]の値を出力し、記憶部に格納する。

【００３７】

一方、ステップＳ４４の判断において変数probe_positionがm以下の場合(no)

には、ステップＳ５２に示す評価処理を実行する。評価処理は、より詳細には後述するが、position[i]の値を決定し、評価結果を記憶部に格納する処理を実行させる。詳細には後述するように評価処理が所定の停止条件を満たすことにより終了すると、処理はステップＳ４６においてmin_kの値が最大編集距離kよりも大きくなったか否かを判断する。min_kがkよりも大きい場合（yes）には、(m-k)以上の共通サブシーケンスが存在しないので、ステップＳ５０に進んで、(n+1)をpositions[i]の値として出力させ、評価処理を停止させる。ステップＳ４６の判断における“true”の値が評価停止を行うための信号として生成される。また、ステップＳ４６の判断において、min_kの値がkより大きくない場合(no)には、その後処理は、ステップＳ５４へと進み、probe_positionの値をインクリメントして、処理をステップＳ４４へと戻し、相補配列Ｑにおける次の塩基の評価を再度実行させる。ステップＳ４４からステップＳ５２の処理は、min_kの値を動的に変化させることで評価処理の停止条件を動的に変更させながら、ステップS４６の判断が肯定的な結果(yes)を返すまで繰り返される。

【００３８】

図６は、図５のステップＳ４０に示した初期化処理の処理を示した図である。図６において示した初期化処理は、ステップＳ６０において編集距離カウンタiを０に初期化し、ステップＳ６２およびステップＳ６４において、編集距離カウンタiが最大編集距離kよりも大きくなるまでposition[i]=0としてセットする。同時に編集距離カウンタiを、i=i+1へとインクリメントする。この初期化処理は、編集距離k以下に対し、positions[i]の値が確実に０となるように設定するために使用される。

【００３９】

図７は、図５において説明したステップＳ５２の評価処理の詳細なプロセスを示したフローチャートである。図７に示した評価処理は、ステップＳ７０においてmax_kを最大編集距離の値k、またはprobe_positionの値のうち、小さい方の値として設定する。ステップＳ７２では、編集距離カウンタiが停止条件であるmin_kよりも小さいか否かを判断する。編集距離カウンタiの値がmin_k以上の場合には(no)には、ステップＳ７４に進んで編集距離カウンタiが０か否かを判断する

。また、ステップＳ７２の判断において編集距離カウンタiがmin_kより小さいと判断される場合(yes)には、評価処理を停止し、図５のステップＳ５４へと分岐する。

【００４０】

ステップＳ７４の判断において、編集距離カウンタiが０の場合(yes)には、ステップＳ７６へと進んでpositions[i]の値をnext_char_position(T, positions[i]+1,Q(probe_position))として与える。

【００４１】

一方、ステップＳ７４の判断において、i=0ではないと判断された場合(no)には、処理は、ステップＳ７８へと分岐し、positions[i]の値を、next_char_position
(T, positions[i]+1,Q(probe_position))の値、またはpositions[i-1]の値のいずれか小さい方の値として設定する。したがって、positions[i]がその時点で評価している塩基数n以下の値を返す場合には、所定の最大編集距離k内において一致する塩基配列が存在する可能性があることを意味する。ステップＳ７６およびステップＳ７８において算出されたpositions[i]の値は、評価結果として適切な記憶部に格納される。

【００４２】

その後、処理はステップＳ８０へと進んで、positions[i]がnよりも大きくなったか否かを判断する。上述したように、関数next_char_position（）は、評価している塩基に該当するものが見出されない場合であてもその値として相補配列Qの塩基数であるn以下の値を返す。したがって、ステップＳ８０の判断において、positions[i]の値がm以下である場合(no)には、さらに適合する可能性を判断するために、ステップＳ８４において編集距離カウンタiを、i=i-1として降順に設定し、ステップＳ７２へと処理を戻してさらに判断を続行させる。また、ステップＳ８０の判断が(yes)の場合には、図５に示したlargest_prefix_with_common_
subsequence()によりpositions[i]の値がn+1として与えられ、図７の評価処理を終了させるべく、ステップS８２においてmin_k=i+1として設定した後、ステップ

Ｓ８４へと処理を戻す。ステップＳ８４では、i=i-1に設定した後、処理をステップＳ７２へと戻すので、ステップＳ７２において確実に判断結果として真の値(yes)を返し、図５に示したステップＳ５４へと分岐させ、処理を繰り返す。

【００４３】

図８には、本発明のスクリーニング方法において、ターゲット塩基配列の評価を部分配列ごとに実行させる実施の形態を示す。図８に示した実施の形態では、ターゲット塩基配列Ｔに対して部分配列の所定の長さを規定する変数lengthと、部分配列のターゲット塩基配列における位置を指定するpositionと、部分配列間の末端部における重なり合いを与える変数であるoverlap_lengthとを定義する。図８に示す本発明のスクリーニング方法の処理は、ターゲット塩基配列の塩基位置を、変数position、変数length、変数overlap_lengthを使用して部分配列ごとに評価をすることを除き、図５～図７に示した処理を適用することができるので、本発明の部分配列ごとに区切って評価を行う処理の本質的な部分のみを説明する。ステップＳ９０において、まず部分配列を指定する変数positionを初期設定する。次いでステップＳ９２において変数positionがｎ以下であるか否かを判断する。ステップＳ９２の判断では、ターゲット塩基配列の最末端であるｎまで評価が実行されるまではpositionはｎ以下なので、ステップＳ９２の判断は、真(yes)とされ、ステップＳ９４へと進み変数lengthを定義する。この変数lengthは、あらかじめ実験条件などを考慮して設定またはユーザ入力により定めることができるし、本発明においては、他の適切な関数、例えば図５に示したlargest_prefix_length
_with_common_subsequence()の返り値を使用して、ステップＳ９４に示すようにして定めることもできる。

【００４４】

関数largest_prefix_length_with_common_subsequence()は、上述したようにランダムで充分に長いターゲット塩基配列と、ランダムな相補配列Ｑとが与えられた場合、4mの値を返す関数である。本発明においては上述した特定の関数を使用するのではなく、適切な値を取るようにlengthを設定することができる。また、図８に示した実施の形態においても編集距離カウンタiについては、指定され

た部分配列ごとに初期化設定を行い、ターゲット塩基配列Ｔとして部分配列T_wを使用することにより、図５～図７に示した処理を繰り返して実行させることができる。ステップＳ９６では、変数lengthの値が、塩基相互の距離にかかわらず結合する可能性がある程度高いことが想定される塩基鎖長であるcheck_lengthよりも短いか否かを判断する。

【００４５】

ステップＳ９６の判断において、lengthがcheck_lengthよりも短い場合(yes)には、ステップＳ１０４へと進んでSmith-Waterman法(P.D. Smith, M.S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147: 195-197, 1981.)などの高精度な手法を使用して結合性の判断を行ない、check_lengthを加えて新たなposition値を生成して、ステップＳ９２へと処理を戻すこともできる。この際、check_lengthの値をm/r以上の値とすることもできる。また、４mの値として設定すると、部分配列すべてについて「結合する可能性がある」と判断し、ステップＳ１０４の評価を高頻度に実行することになるので、本発明の高速性が阻害されてしまうことになる。このため、本発明においてはcheck_

lengthの値をm/r～4mの間の値として設定することができ、check_lengthの値として概ね2m程度の値を採用することが、安定した結果を与えることが多いことが見出された。

【００４６】

一方、ステップＳ９６の判断においてlengthがcheck_length以上の場合(no)には、ステップＳ９８においてlengthと残りの塩基鎖長とを比較し、lengthが残りの塩基鎖長よりも短い場合(yes)には、さらにステップＳ１０２へと進んで、それまで評価を行った塩基位置にlengthを加え、さらに適切な重なり合いを与えることができるようにoverlap_length分を考慮して次の部分配列を指定する。図８に示した実施の形態では、その後再度ステップＳ９２へと処理を戻し、ステップＳ９２～ステップＳ９８までの処理を繰り返す。また、lengthが残りの塩基鎖長を越えた場合(no)には、ステップＳ１００においてposition=n+1に設定し、処理をステップＳ９２へと戻す。この場合、ステップＳ９２の判断において偽の値が

返され(no)、ステップＳ１０６において結合しないという評価結果が与えられる。

【００４７】

図９には、図５～図７に示した処理を実行させるための擬似コードを示す。また、図１０には、図８に示した部分配列ごとに評価を実行するための処理の擬似コードを示す。図１０において使用された関数check_exactly(T, Q, k)は、ステ図８のステップＳ１０２において使用される高精度の判定手法を示す。関数check_exactly(T, Q, k)は、ターゲット塩基配列Ｔと相補配列Ｑとが実際に結合するかどうかの判定を、従来の時間はかかるものの、正確に計算できるSmith-Watermanなどの方法を使用して計算し、結合しないならばtrueを、結合するならばfalseを返す。本発明においては、check_exactly(T, Q, k)としては、Smith-Watermanの方法を具体的には挙げることができるが、Smith-Watermanの方法以外にでも、いかなる方法を使用しても構成することができる。また、関数check_exactly(T, Q, k)について、結合しないことを高精度にユーザに対して提示するために、この関数を常にFALSEを返す関数にすることもできる。この場合には、「結合しない」という結果が出た場合は必ず結合しない、ということが言えるので、ハイブリダイゼーションの判定にも有用である。

【００４８】

図１１は、本発明のスクリーニング方法により与えられる評価結果をテーブルの形式で示した図である。図１１に示した実施の形態においては、ターゲット塩基配列ＴがCGCGCATGAAであり、相補配列ＱがGCCCATGCであって、編集距離ｋ＝３の場合に得られる実施の形態である。以下、図１１を使用して本発明のスクリーニング方法により得られる評価結果について説明する。図１１に示した表の縦欄は、編集距離カウンタiの値に対応し、横欄は、相補配列中位置を表すprobe_positionの値を示し、表中の値は、図７における評価処理の終了時のposition[i]の値を示している。図１１に示した実施の形態では、最大編集距離k=3なので、編集距離カウンタiが3以下のものについてのみ計算が行われている。図１１に示した実施の形態では、positions[i, probe_position]の計算は、最終的に図１１に示した擬似コードの②まで計算が進み、positions[3]の値として7が得られてい

るため、返り値として７を返し、このときmin_kの最終的な値は、2となる。上述したpositions[i, probe_posiiton]は、positions[i, probe_posiiton]が、jの時のpositions[i]の値を、probe_position[i, j]とする。この配列データは、テーブルとして格納しておくこともできるし、本発明において説明するように、評価を行っているprobe_positionの変数に対するposition[i]の値のみを格納させておいても良い。

【００４９】

図１２（ａ）、（ｂ）には、図１１に示したpositions[0]およびpositions[1]についての値を返す処理を図式的に示す。図１２（ａ）に示されるように、positions[0]の場合には、相補配列Ｑにおいて出現する塩基が単純にターゲット塩基配列において出現する位置となり、2，3，5で与えられる値がpositions[1，m]の値として与えられている

【００５０】

一方でpositions[1]についてみれば、i≧2については、next_char_positionと、positions[i-1]の値のうちの小さい方の値が返り値とされるので、positions[0，1]の値は２であるものの、相補配列Ｑのシトシン（Ｃ）がターゲット塩基配列の１番目に出現しているので小さい方の１が得られている。同様に相補配列Ｑの２番目のシトシン（Ｃ）は、ターゲット塩基配列Ｔの４番目に出現しており、next_char_positionの返り値は4となるものの、これは対応するpositions[0，2]の値が3のため、小さい方の値3が返されている。

【００５１】

さらに、相補配列Ｑの４番目の塩基であるシトシン（Ｃ）は、５番目に出現し、アデニン（Ａ）は、６番目に出現し、チミン（Ｔ）は、７番目に出現し、７番目のグアニン（Ｇ）は、８番目に出現しており、これらの番号がpositions[1]の行を構成し、1，3，5，6，7，8の行が与えられることになる。同様にして、position[2]については、相補配列Ｑの３番目のシトシン（Ｃ）がターゲット塩基配列Ｔの１番目に出現し１の値を与え、４番目のシトシン（Ｃ）が３番目に出現し、５番目のアデニン（Ａ）が６番目に出現しているが、対応するpositions[1]のm=4の値である５が採用され、６番目のチミン（Ｔ）が７番目に出現しているも

ののpositions[1]のm=5の値である６が採用されるので図１１に示されるように、0，1，3，5，6，7，8の列が与えられているのが示される。

【００５２】

図１１に示された実施の形態であるk=3の場合、positions[1]の値がprobe_position=7の地点で8の値をとる。そのため、next_char_position（T，9，“C”の値がn+1=11となり、nよりも大きな値なので、最終的なprobe_position=8の処理の際には、min_kの値が１だけインクリメントされ、2が与えられる。このmin_k=2の値は、ターゲット塩基配列および相補配列に共通するサブシーケンスの長さがm-2=6である、ということを示している。また、同じターゲット塩基配列と相補配列とを使用し、最大編集距離k=1として同様の計算を実行させると、編集距離カウンタi=0,1についてのpositions[1]およびpositions[0]の計算が実行され、評価結果が得られ処理を終了する。この場合、擬似コードの①まで計算が進み、n+1=11という値を返す。これは、m-1=7以上の長さの共通するサブシーケンスが存在しないことを示す。本発明においては、図１１に示した評価結果テーブルとして必ずしも評価結果を蓄積する必要はなく、上述したように最大のpositions[i]の値についてだけ格納させるデータ形式で記憶部に評価結果を格納しておくこともできる。

【００５３】

Ｃ．本発明の塩基配列スクリーニング・システム

図１３は、本発明のスクリーニング方法を実装した塩基配列スクリーニング・システムの該略的な機能ブロック図である。図１３に示された塩基配列スクリーニング・システム（以下、システムと略する。）１０は、概ね中央処理装置（ＣＰＵ）および必要な記憶装置とを含むコンピュータを含んで構成されており、ターゲット塩基配列や相補配列の入力を受けて、スクリーニングのための評価処理を実行させている。

【００５４】

さらに詳細に説明すると、本発明のシステム１０は、相補配列格納部１４と、ターゲット塩基配列格納部１６と、最大編集距離格納部１８と、評価処理部２０とを含んで構成されている。相補配列格納部１４、ターゲット塩基配列格納部１

６および最大編集距離格納部１８は、ぞれぞれ適切なメモリまたはメモリ領域として構成されている。図１３においては、機能ブロックを明確に表示させるべく、異なるブロックとして記述されているものの、本発明においては、実装上の都合により、単一のランダム・アクセス・メモリなどに割り当てられたメモリ領域として構成することもできる。相補配列格納部１４は、入力部１２からのプローブＰの入力を受け取り、相補性の関係を使用して生成された相補配列Ｑを格納する。また、ターゲット塩基配列格納部１６および最大編集距離格納部は、入力されたターゲット塩基配列のデータおよび最大編集距離ｋを、それぞれ格納している。上述した入力部１２は、本発明においては、キーボード、データ読み取り装置などを含んで構成することができ、データ読み取り装置は、例えばフロッピー（登録商標）・ディスク、ハードディスク、ＣＤ－ＲＯＭ、ＤＶＤといった入出力媒体などに記憶されたプローブＰの塩基配列を読み出し、本発明のシステムへとプローブＰのデータを送っている。

【００５５】

評価処理部２０は、本発明のスクリーニング方法を実行させるためのプログラムを、図示しないハードディスクなどからＣＰＵが読み出して、ソフトウエア的に構成される機能ブロックとして実装される。評価処理部２０は、図５に示したlargest_prefix_with_common_subsequenceにより構成され、動的に評価を停止させる停止判断部２０ａと、評価実行部２０ｂとを含んで構成されている。評価処理部２０は、相補配列格納部１４、ターゲット塩基配列格納部１６、および最大編集距離格納部１６から、それぞれ相補配列Ｑ、ターゲット塩基配列Ｔ、最大編集距離ｋを読み出し、本発明のスクリーニング方法を適用して、共通するサブシーケンスの評価・判断を実行する。停止判断部２０ａは、min_kの値と最大編集距離kとを比較して、min_kの値がkよりも大きくなった場合には、ユーザの要求する範囲内において「結合する」と判断されるシーケンスがないものと判断して、評価実行部２０ｂに対してそれ以後のprobe_positionについて評価を行わないように指令する。この指令を受け取った評価処理部２０は、計算された評価結果を、本発明における特定の実施の形態では、positions[i, probe_position]のフォーマットとして記憶部２２へと書き込みを行なう。本発明の特定の実施の形態

においては、評価結果は、図１１に示した評価結果テーブルとして構成することが後のさらなる評価のための情報を与える点では好ましいが、本発明においてはまた、必要とされるpositions[i, probe_position]の値だけを選択して格納することもできる。

【００５６】

評価処理部２０は、評価結果テーブルの作成を終了すると、positions[i]の値を参照し、所定のprobe_positionの値に対応するposition[i]の値をルックアップし、probe_positionに対応するコラムにおいてprobe_position以上の返り値が含まれているか否かを判断する。probe_position＞mで、かつmin_kの値がkよりも大きくなった場合には、要求された最大編集距離kに対応する範囲内で共通するサブシーケンスが見出されないと判断し、「結合しない」という判断結果をユーザに対して表示部２４において表示させる。また、probe_position＞mであって、min_k＜kである場合には、当該positions[i]の編集距離カウンタiにおいて、min_k=i+1であり、シーケンス長が(m-min_k)で与えられる共通するシーケンスがあるものと判断される。この判断結果は、表示部２４を通してユーザに「結合する」画面表示と共に、シーケンスの塩基配列およびシーケンス長といった付帯的情報を表示させることにより、提供される。ユーザに対して提供された判断結果は、ユーザの必要性に応じて、プリンタからハードコピーされるか、またはフロッピー（登録商標）・ディスク、ハードディスク、ＣＤ－ＲＯＭ、ＤＶＤ、フラッシュメモリなどの記憶媒体に格納することができる。

【００５７】

図１４は、本発明のスクリーニング方法の第２の実施の形態を実装したシステムの実施の形態を示した図である。図１４に示したシステム１０は、評価処理部２０の構成を除き、概ね図１３において説明したと同一の構成とされている。このため、図１４に示したシステム１０について、評価処理部２０について詳細に説明する。図１４に示した評価処理部２０は、停止判断部２０ａと、評価実行部２０ｂと、部分配列指定部２０ｃとを含んで構成されている。部分配列指定部２０ｃは、まず最初に評価を開始する部分配列を、指定データ格納部２８からpositionの値およびlengthの値などを読み出して決定する。指定された部分配列は、

評価実行部２０ｂへと渡されて、図１３において説明したと同様に評価が実行される。停止判断部２０ａは、評価実行部２０ｂへと評価の停止を指令すると共に、部分配列指定部２０ｃに対しても部分配列について評価が終了したことを指令し、さらに評価を行うべき部分配列がある場合には、次に評価する部分配列を指定して、評価実行部２０ｂに部分配列のデータを転送する。

【００５８】

図１５は、本発明のシステム１０の第２の実施の形態を示した図である。図１４に示した本発明の第２の実施の形態のシステムは、ネットワーク３０を介して、ターゲット塩基配列Ｔおよびプローブ Ｐの塩基配列のデータをユーザから受信する。受信されたデータは、送受信部３４の入力バッファに格納された後、相補配列格納部１４、ターゲット塩基配列格納部１６、最大編集距離格納部１８、および指定データ格納部２８に格納される。この場合、ユーザは直接ターゲット塩基配列Ｔを送信することもできるが、ｃＤＮＡデータベースなど、システム１０がターゲット塩基配列Ｔについての情報を格納したデータベースを含んで構成される場合には、評価処理部２０は、ネットワークを介してユーザから受け取ったターゲット指定データを参照して、データベース３２にアクセスして該当するターゲット塩基配列を取得し、取得したターゲット塩基配列Ｔをターゲット塩基配列格納部１６に格納する。評価処理部２０は、図１４において説明したと同様の処理を実行して評価結果を生成し、記憶部２２に格納すると共に、評価結果を送受信部３４へと送り、ネットワーク３０を介してユーザに評価結果を送出して、塩基配列データベースに格納された塩基配列をスクリーニングするためのプローブＰの有効性についての情報をユーザに返す構成とされている。

【００５９】

【実施例】

本発明のスクリーニング方法の効果を検証するため、Windows（登録商標）2000を搭載したパーソナル・コンピュータ（Intel社製のPetinum（登録商標）4：2.2GHzプロセッサを実装）に本発明のプログラムを実装して、従来の高精度法(Smith-Waterman法)と対比して計算速度の評価検討を行った。使用したターゲット塩基配列は、ライブラリから入手した4,639,221bpの塩基鎖長の大腸菌のゲノム(Ge

n Bank Accession番号:NC_000913)を使用した。実験用のパーソナル・コンピュータには、部分配列ごとに評価を実行する図９および図１０で示した擬似コードのアルゴリズムを実装した。なお、check_exactlyの関数としては、Smith-Waterman法を使用した。

【００６０】

入力する相補配列Ｑは、大腸菌の上述したゲノムの一部を使用して塩基を削除することにより塩基鎖長が約580bp程度の塩基鎖を生成し、複数のプローブＰとした。最小のウィンドウサイズ(check_length)をプローブの長さの約２倍に相当する1158と固定した。ハイブリダイゼーションすることができる塩基鎖の相似度を、編集距離ｋを設定して、相似度を0.80、0.85、0.90、0.95となるように設定し、相補配列の相似度の異なる４回の実験を行った。また、比較のため、従来技術のSmith-Waterman法の計算時間も計測した。その結果を図１６に示す。

【００６１】

図１６に示された結果によれば、本発明の方法によるスクリーニングの計算速度は、約２倍～約２０倍であることが確認された。計算速度は、相似度を高く設定できる場合は、最大編集距離kを小さく設定するので、プローブＰの塩基配列についての反復計算を制限することができ（すなわち、編集距離カウンタの反復回数がk以下となる）、判断を行う塩基の絶対数が減少することから高速化が可能であることが示された。すなわち、本発明によれば、従来手法とは異なり、相似度を高くすればするほど高精度な計算が可能となることがわかる。

【００６２】

図１７には、図１６に示された表の結果を、縦軸を計算時間（ｓ）、横軸を相似度（％）としてプロットしたグラフ図である。図１７に示されるように、相似度が高くなればなるほど判断結果を得るための時間が短縮することが示されている。また、相似度１００％における判断結果を得るための時間は、比較例であるSmith-Waterman法の結果に比べて少なくとも数１０倍改善されているのが示された。

【００６３】

以上、本発明について、図面に示した具体的な実施の形態をもって説明してき

たが、本発明の上述した各機能を実現する手段または部分は、コンピュータ実行可能なプログラム言語により記述されたソフトウエアまたはソフトウエア・モジュール群として構成することができ、必ずしも図面に記載した機能ブロックとして構成される必要はない。また、本発明の塩基配列スクリーニング・システムでは、各機能要素は、必要に応じていかなる機能モジュールと共に構成することができ、図面に示された特定の実施の形態に限定されるものではない。

【００６４】

さらに、本発明のプログラムは、種々のプログラミング言語、例えばFORTRAN、C言語、C++言語、JAVA（登録商標）などを使用して記述することができ、本発明のプログラムを記述したコードは、磁気テープ、フレキシブル・ディスク、ハードディスク、コンパクト・ディスク（ＣＤ）、光磁気ディスク、デジタル・バーサタイル・ディスク（ＤＶＤ）といったコンピュータ可読な記録媒体に保持させることができるし、伝送媒体としても頒布することができる。

【００６５】

本発明によれば、少なくとも相補性の低いプローブを選択的に排除することにより、従来の高精度なアラインメントによる手法よりも高速に健全性の高いプローブに関する情報をユーザに提供することが可能となる。また、本発明の方法をプローブの１次スクリーニングとして適用し、スクリーニング後に従来の高精度なアラインメントによる手法と組み合わせることで、プローブのスクリーニング処理を、より高速化することが可能な塩基配列をスクリーニングするためのシステム、そのための方法、該方法をコンピュータに対して実行させるためのプログラム、該プログラムを記憶したコンピュータ可読な記録媒体およびネットワークを介して塩基配列のスクリーニングを実行させるためのサーバを提供することができる。

【図面の簡単な説明】

【図１】　本発明におけるターゲット塩基配列と、プローブと、相補配列との関係を示した図。

【図２】　本発明におけるターゲット塩基配列のスクリーニングを行う実施の形態の概略図。

【図３】　本発明のスクリーニング方法の第１の実施の形態の処理を示したフローチャート。

【図４】　本発明のスクリーニング方法の第２の実施の形態の処理を示したフローチャート。

【図５】　本発明において結合性の評価を行う処理の概略的なフローチャート。

【図６】　本発明において結合性の評価を行う処理のための初期化処理を示した図。

【図７】　本発明において結合性の評価のための評価処理の詳細を示したフローチャート。

【図８】　本発明のスクリーニング方法の第２の実施の形態の詳細な処理を示したフローチャート。

【図９】　本発明の結合性の評価を行うための処理を実行させるプログラムの擬似コード。

【図１０】　本発明の第２の実施の形態において部分配列を指定して結合性を評価するためのプログラムの擬似コード。

【図１１】　本発明における評価結果の出力データ構造の実施の形態を示した図。

【図１２】　本発明における評価結果を得るための図９に示した擬似コードの実行する処理を示した図。

【図１３】　本発明の塩基配列スクリーニング・システムの第１の実施の形態を示した図。

【図１４】　本発明の塩基配列スクリーニング・システムの第２の実施の形態を示した図。

【図１５】　本発明の塩基配列スクリーニング・システムの第３の実施の形態を示した図。

【図１６】　本発明の実施例の結果を比較例の結果と共に示した図。

【図１７】　本発明の実施例および比較例の結果をプロットして、評価結果を得るまでに必要な時間を、相似度１００％において比較した図。

【図１８】　ターゲット塩基配列と、プローブとの関係を示した図。

【符号の説明】

１０…塩基配列スクリーニング・システム

１２…入力部

１４…相補配列格納部

１６…ターゲット塩基配列格納部

１８…最大編集距離格納部

２０…評価処理部

２２…記憶部

２４…表示部

２６…出力部

２８…指定データ格納部

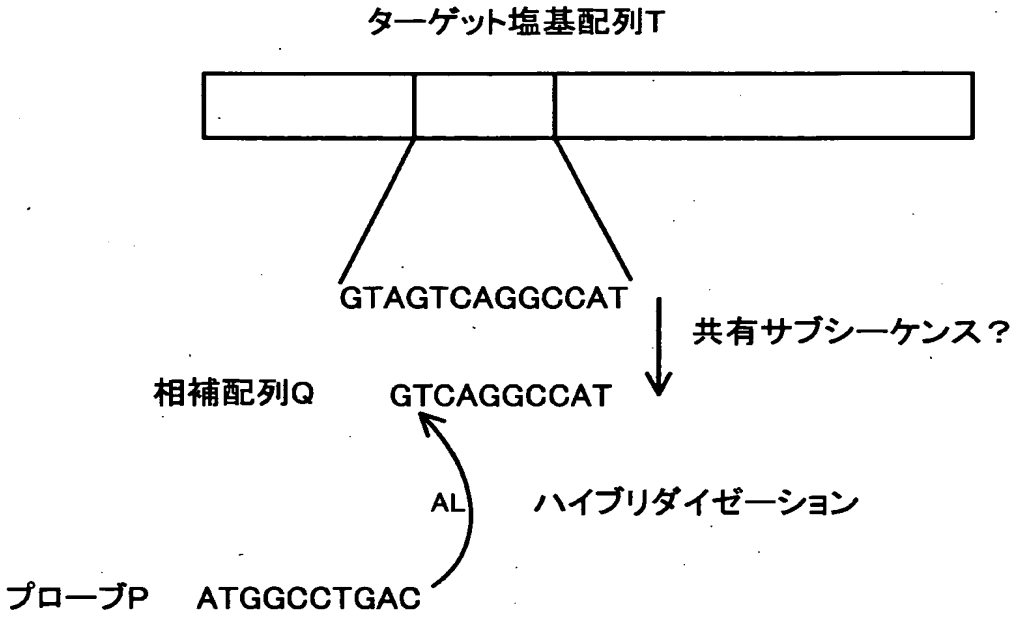
３０…ネットワーク

３２…データベース

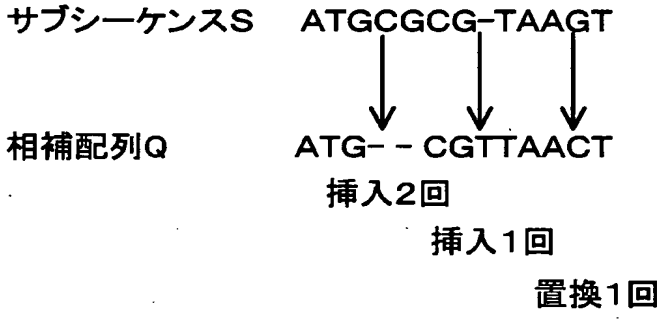
３４…送受信部

【書類名】　図面

【図１】

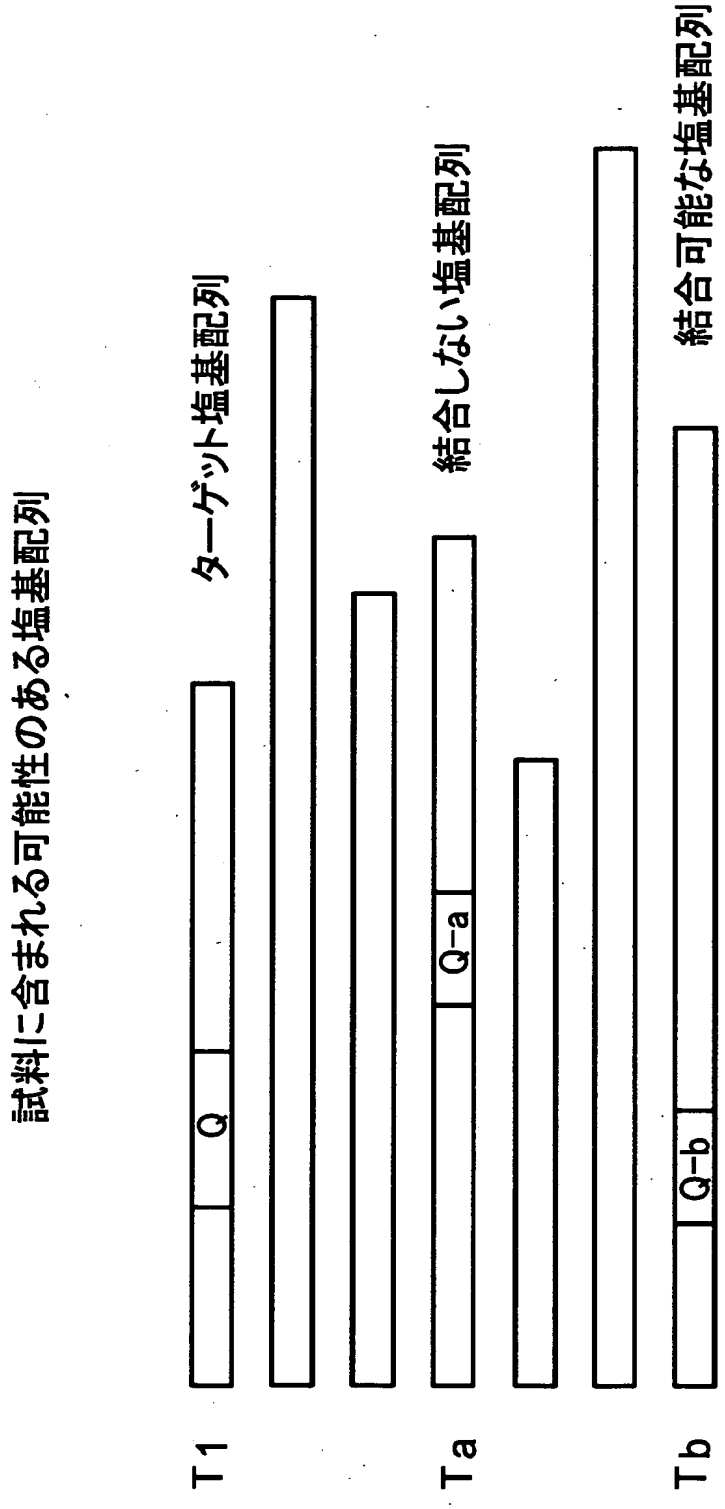

ターゲット塩基配列T
GTAGTCAGGCCAT
共有サブシーケンス？
相補配列Q
GTCAGGCCAT
AL
ハイブリダイゼーション
プローブP
ATGGCCTGAC

(a)

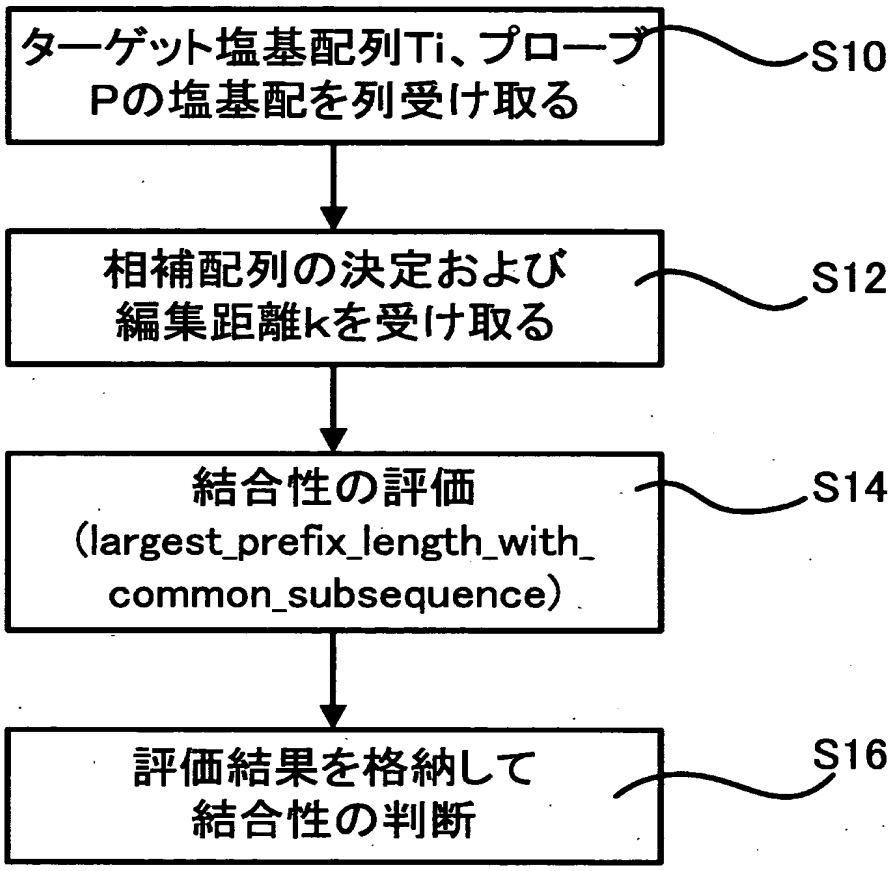

サブシーケンスS
ATGCGCG-TAAGT
相補配列Q
ATG--CGTTAACT
挿入2回
挿入1回
置換1回
相似度r=0.6364
最大編集距離k=4

(b)

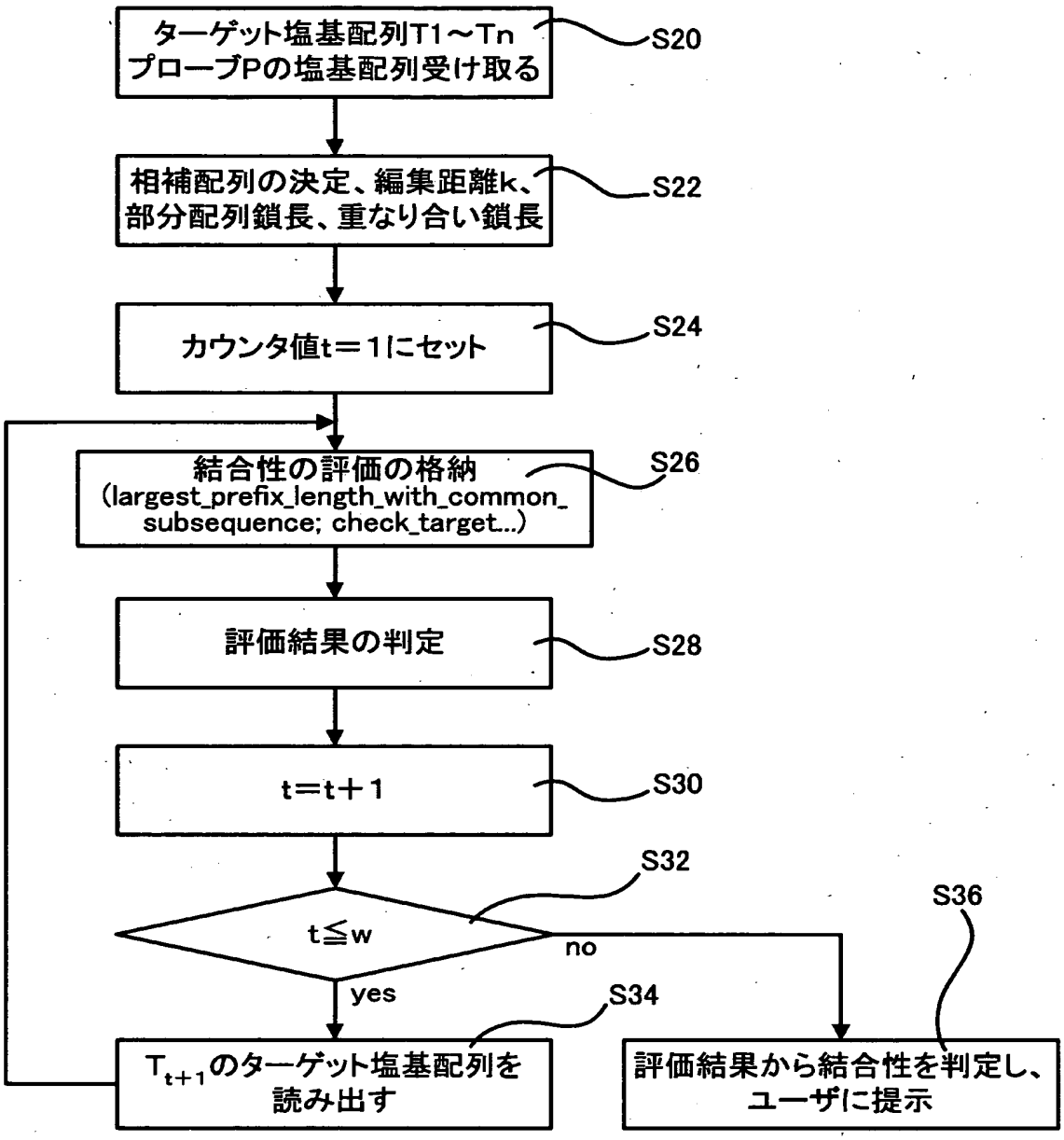
【図２】


試料に含まれる可能性のある塩基配列
ターゲット塩基配列
結合しない塩基配列
結合可能な塩基配列
Q
Q-a
Q-b
T1
Ta
Tb

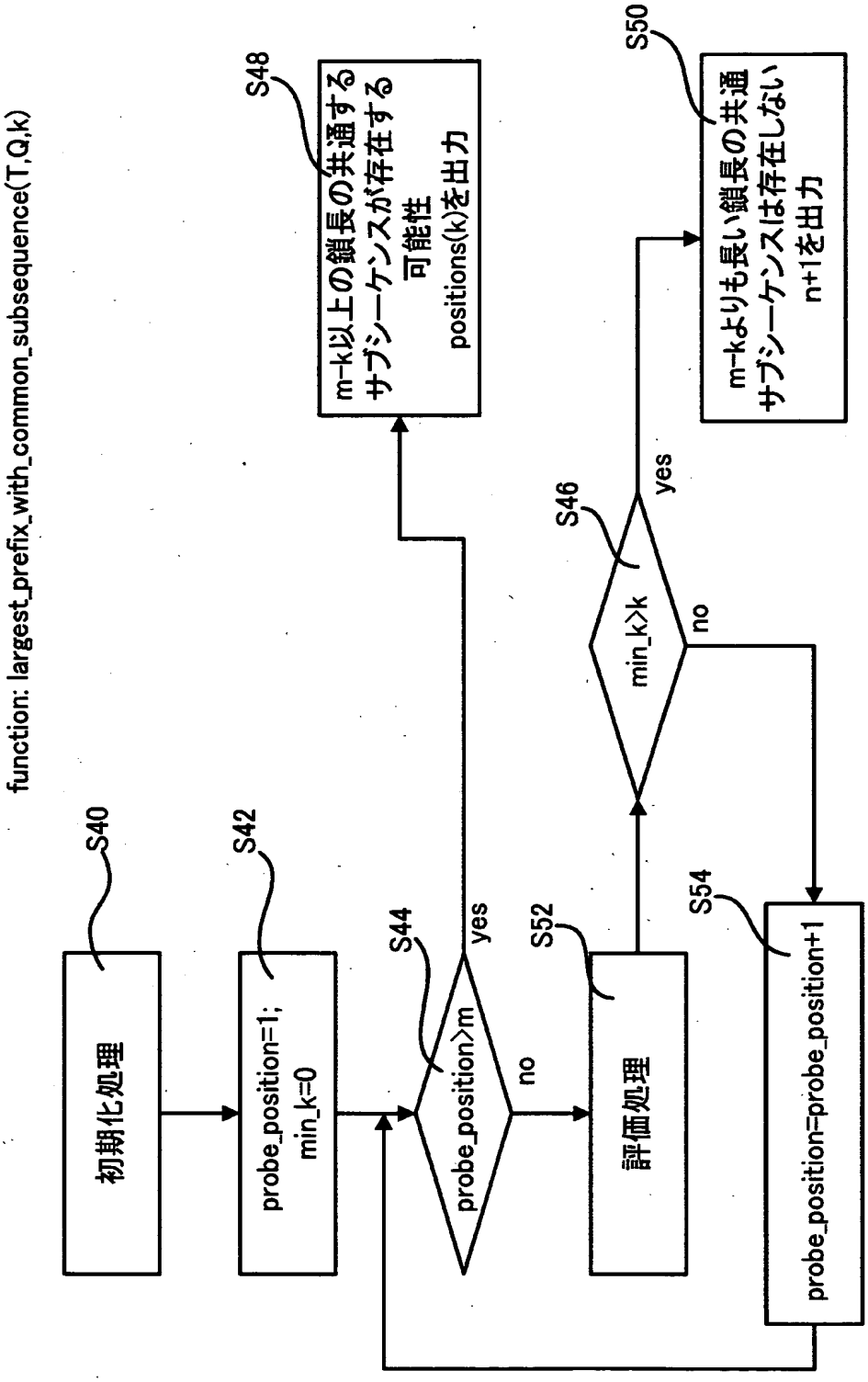
【図３】


ターゲット塩基配列Ti、プローブPの塩基配を列受け取る
S10
相補配列の決定および編集距離kを受け取る
S12
結合性の評価
(largest_prefix_length_with_common_subsequence)
S14
評価結果を格納して結合性の判断
S16

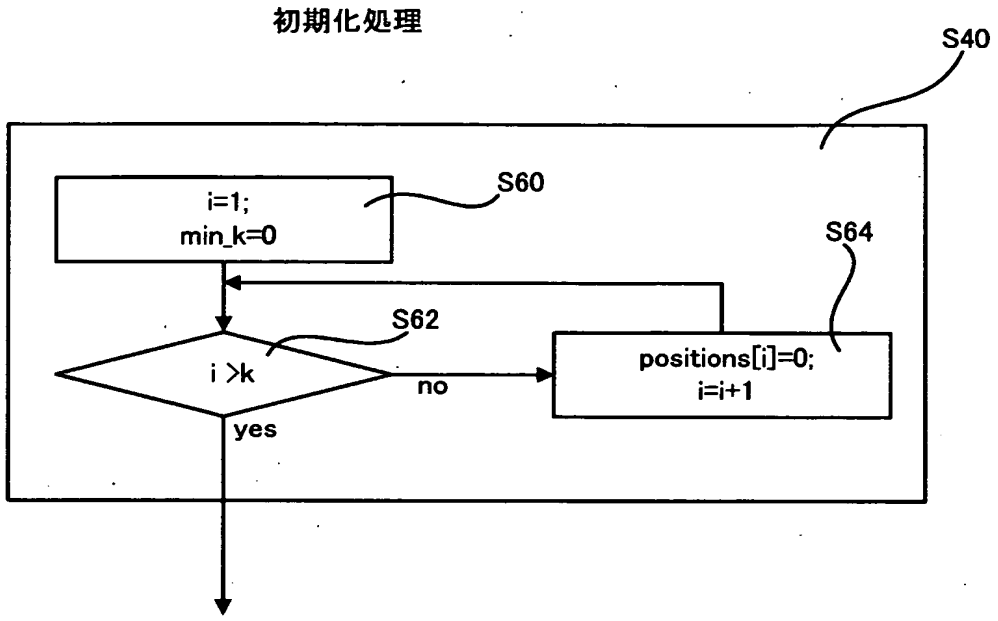
【図４】


ターゲット塩基配列T1～Tn
プローブPの塩基配列受け取る
S20
相補配列の決定、編集距離k、
部分配列鎖長、重なり合い鎖長
S22
カウンタ値t＝1にセット
S24
結合性の評価の格納
(largest_prefix_length_with_common_
subsequence; check_target...)
S26
評価結果の判定
S28
t＝t＋1
S30
S32
t≦w
no
yes
S34
S36
T_{t+1}のターゲット塩基配列を
読み出す
評価結果から結合性を判定し、
ユーザに提示

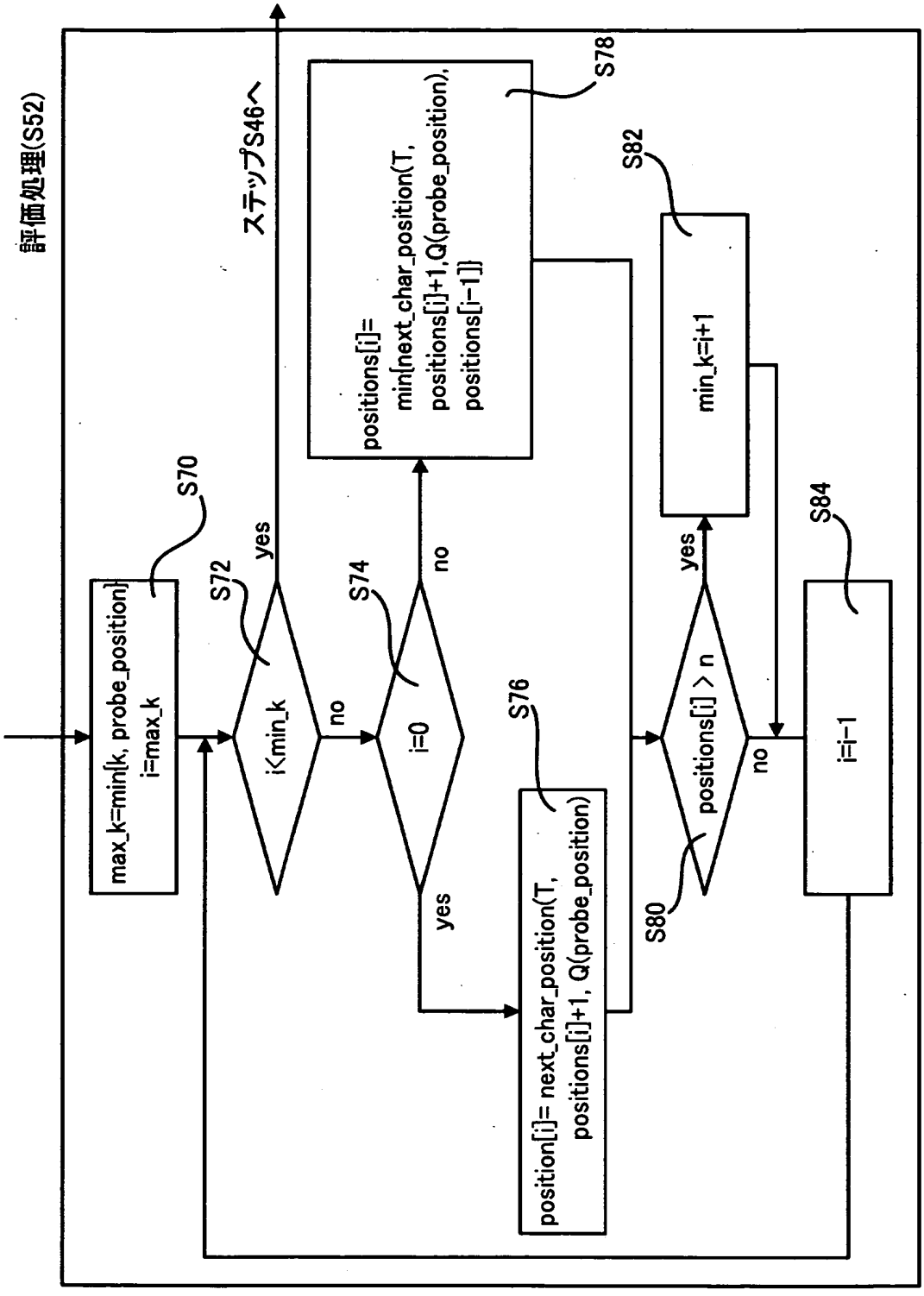
【図５】


function: largest_prefix_with_common_subsequence(T,Q,k)
初期化処理
S40
probe_position=1;
min_k=0
S42
probe_position>m
S44
yes
no
評価処理
S52
min_k>k
S46
yes
no
m−k以上の鎖長の共通するサブシーケンスが存在する可能性positions(k)を出力
S48
m−kよりも長い鎖長の共通サブシーケンスは存在しないn+1を出力
S50
probe_position=probe_position+1
S54

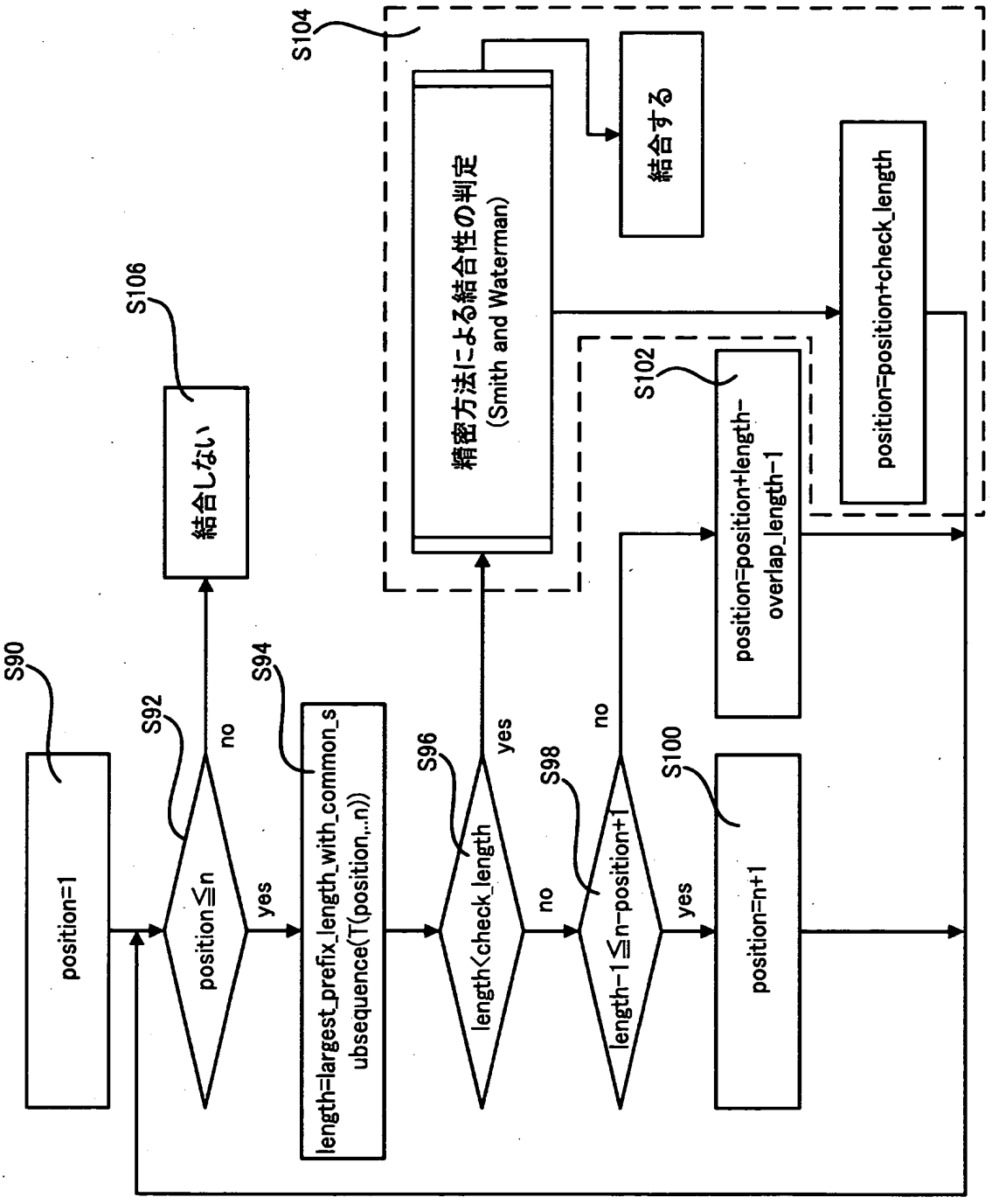
【図６】



【図７】


評価処理(S52)
ステップS46へ
S70
max_k=min[k, probe_position]
i=max_k
S72
i<min_k
yes
no
S74
i=0
no
yes
S76
position[i]= next_char_position(T,
positions[i]+1, Q(probe_position)
S78
positions[i]=
min[next_char_position(T,
positions[i]+1,Q(probe_position),
positions[i-1]]
S80
positions[i] > n
yes
no
S82
min_k=i+1
S84
i=i-1

【図８】


S104
精密方法による結合性の判定
(Smith and Waterman)
結合する
S106
結合しない
S102
position=position+length-
overlap_length-1
position=position+check_length
S90
position=1
S92
position≦n
no
yes
S94
length=largest_prefix_length_with_common_s
ubsequence(T(position,..n))
S96
length<check_length
yes
no
S98
length-1≦n-position+1
no
yes
S100
position=n+1

【図９】

```
//関数 largest_prefix_with_common_subsequence

largest_prefix_with_common_subsequence(配列T[1..n]、配列Q[1..m]、k) {
    for (i=0;i<=k;i++) positions[i] = 0;
    min_k = 0;
   for (probe_position=1; probe_position<=m; probe_position++) {
      max_k = min {k, probe_position};
      for (i=max_k; i>=min_k ;i--) {
        if (i==0) {
                positions[i] = next_char_position(T, positions[i]+1,
Q[probe_position]);
        } else {
                positions[i] = min { next_char_position(T, positions[i]+1,
Q[probe_position]),
                                positions[i-1] }
        }
        if (positions[i] > n) {
                min_k = i+1;
        }
       }
       if (min_k > k) {
         ① m-k以上の長さの共通のサブシーケンスは存在しない
         return m+1;
       }
   }
   ②  m-min_k以上の長さの共通のサブシーケンスが存在する
   return positions(k)
```

【図１０】

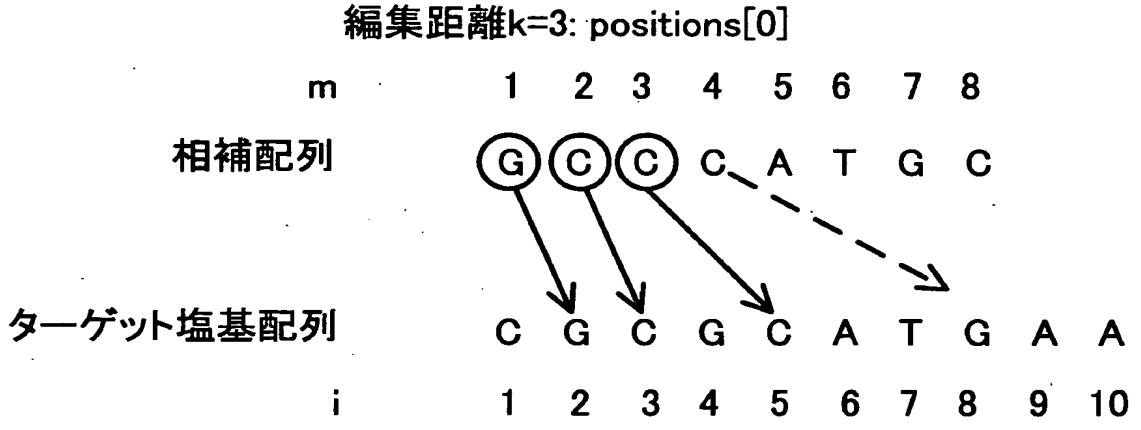
```
//関数 check_target

check_target(配列T[1..n], 配列Q[1..m],k, overlap_length, check_length){
  position=1;
  while(position <= n) {
    length = largest_prefix_length_with_common_subsequence(
                        T[position..n], Q[1..m], k);
    if (length < check_length) {
        if (check_exactly(
                T[position..(min(n, position+check_length))], Q, k)) {
          position = position + check_length;
        } else {
          return ("TとPは結合する")
        }
    } else if (length -1 <= n - position + 1){
      position = position + length - overlap_length - 1;
    } else {
      position = n+1;
    }
  }
 return ("TとPは結合しない");
}
```

【図１１】

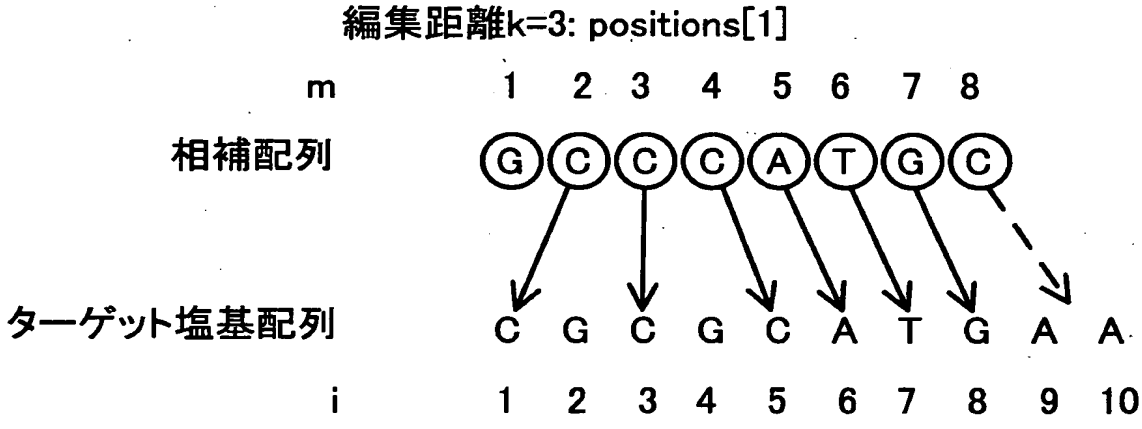
	1	2	3	4	5	6	7	8
positions[0]	2	3	5	−	−	−	−	−
positions[1]	0	1	3	5	6	7	8	−
positions[2]	−	0	1	3	5	6	7	8
positions[3]	−	−	0	1	3	5	6	7

【図１２】



[positions[0],m]= (positions[0], 2, 3, 5, −, −, −, −, −)

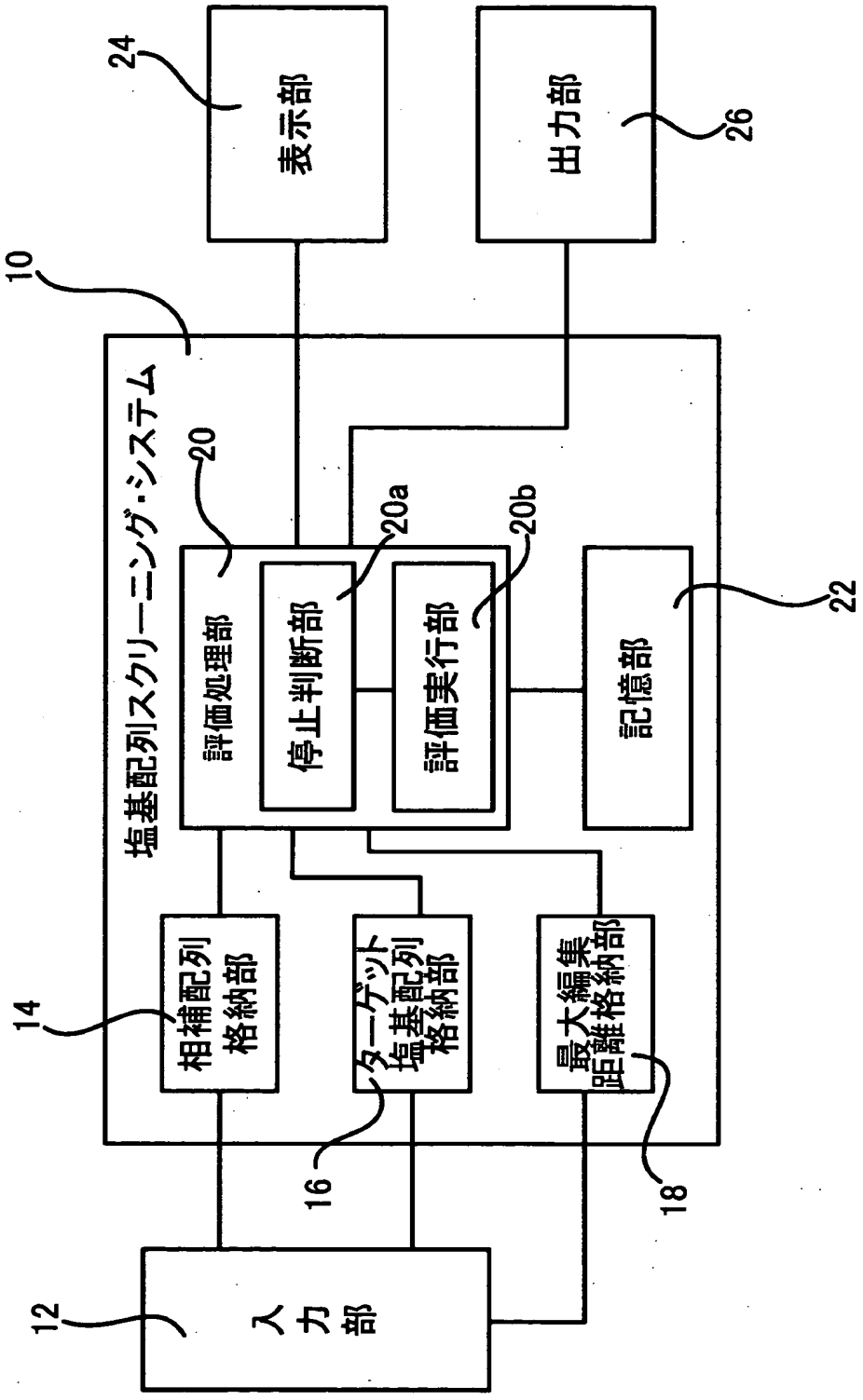
(a)



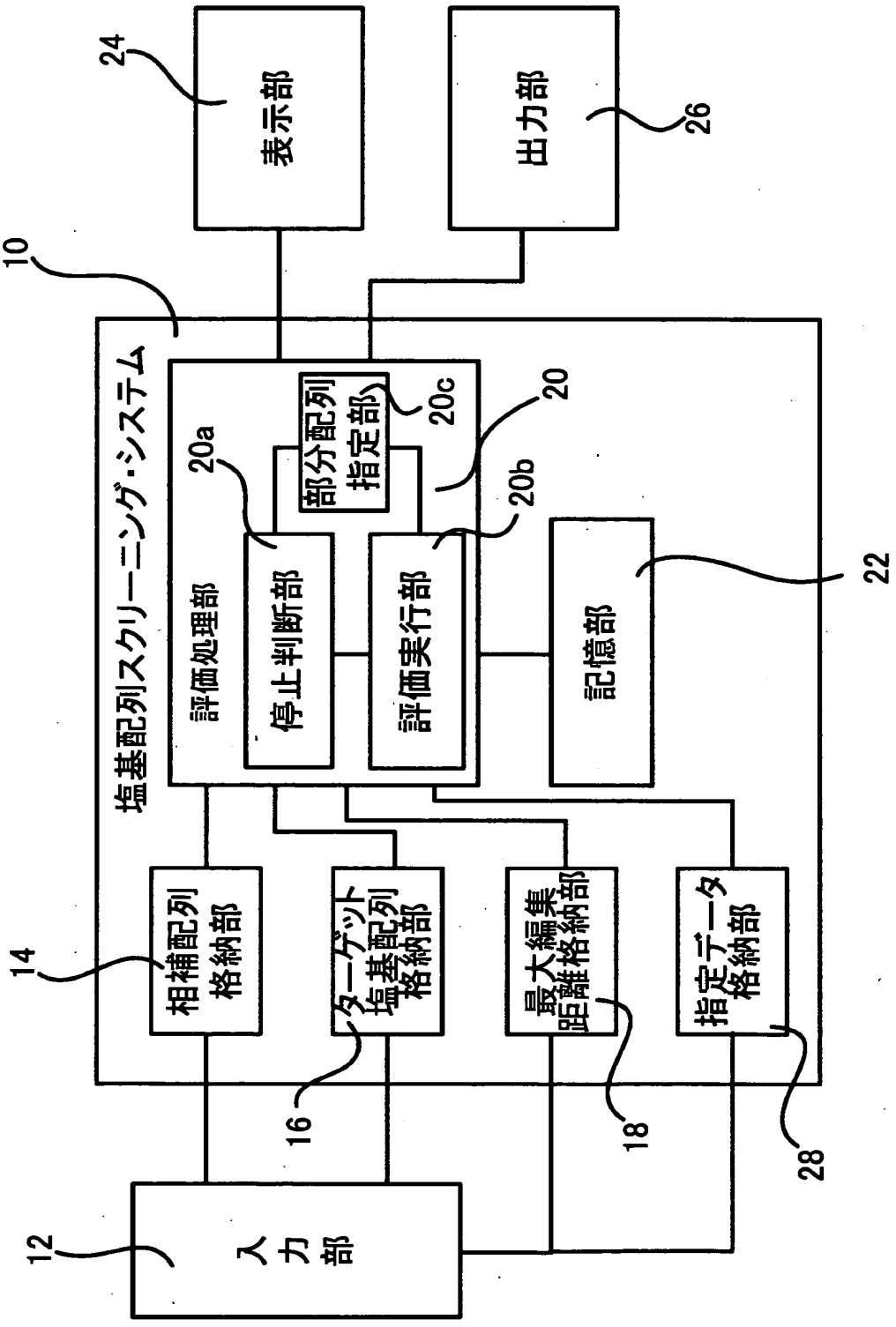
[positions[1],m]= (positions[1], 1, 3, 5, 6, 7, 8)

(b)

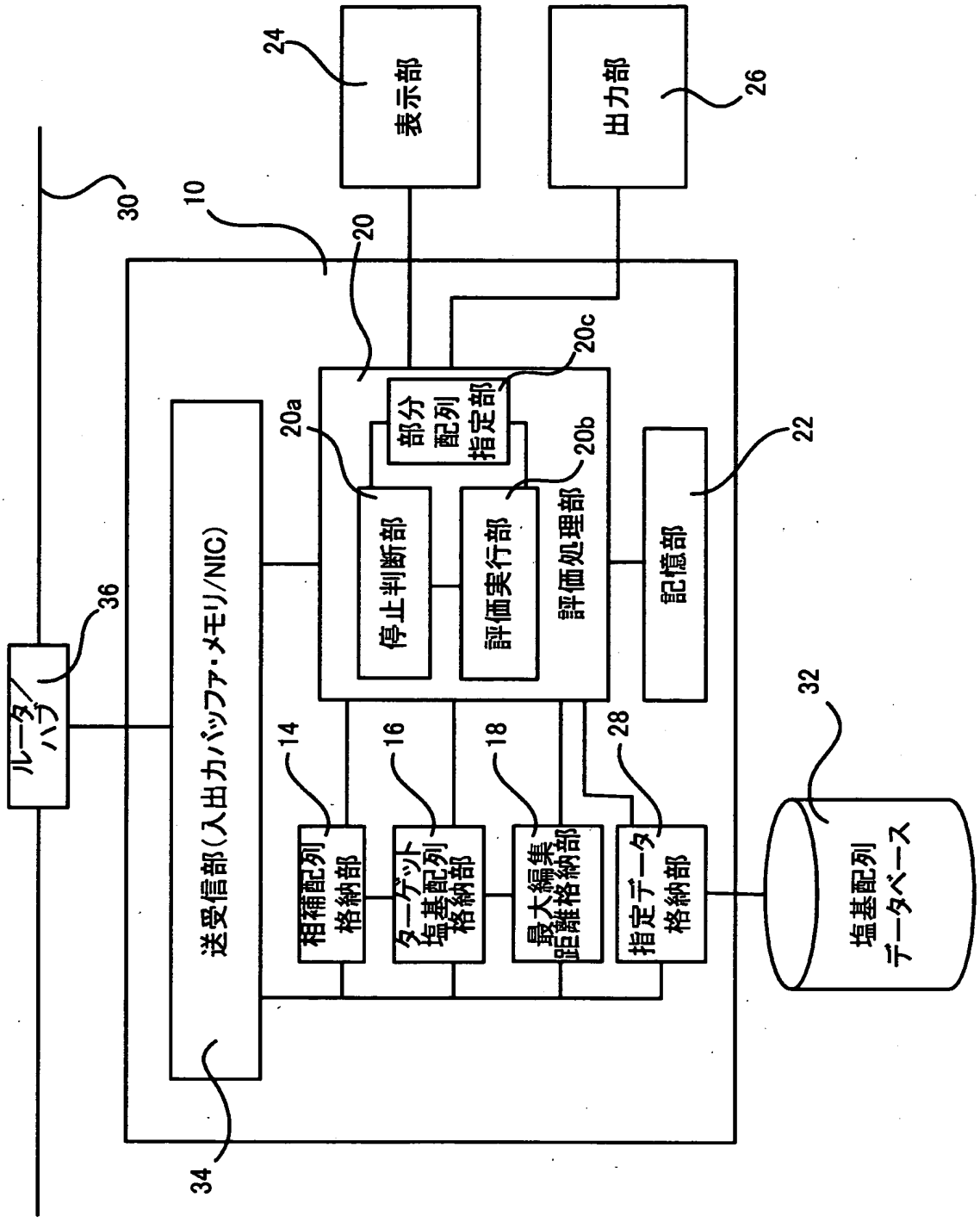
【図１３】


24
表示部
出力部
26
10
塩基配列スクリーニング・システム
20
評価処理部
20a
停止判断部
20b
評価実行部
記憶部
22
14
相補配列格納部
ターゲット塩基配列格納部
16
最大編集距離格納部
18
12
入力部

【図１４】


表示部
24
出力部
26
10
塩基配列スクリーニング・システム
評価処理部
20a
停止判断部
部分配列指定部
20c
20
20b
評価実行部
記憶部
22
相補配列格納部
14
ターゲット塩基配列格納部
16
最大編集距離格納部
18
指定データ格納部
28
入力部
12

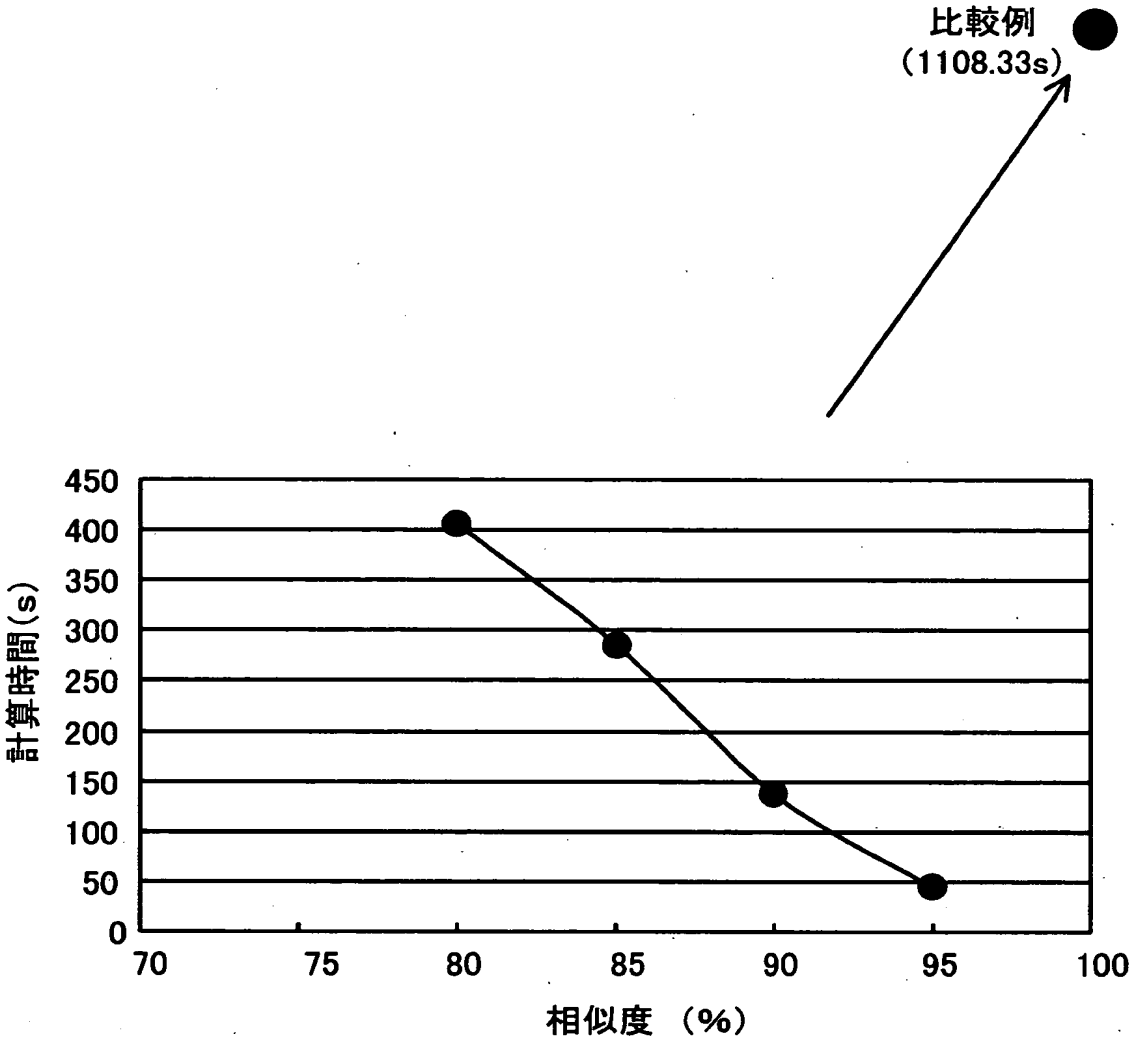
【図１５】


24
表示部
出力部
26
30
10
20
20c
部分配列指定部
20a
20b
22
停止判断部
評価実行部
評価処理部
記憶部
ルータ/ハブ
36
送受信部(入出力バッファ・メモリ/NIC)
32
14
16
18
28
相補配列格納部
ターゲット塩基配列格納部
最大編集距離格納部
指定データ格納部
塩基配列データベース
34

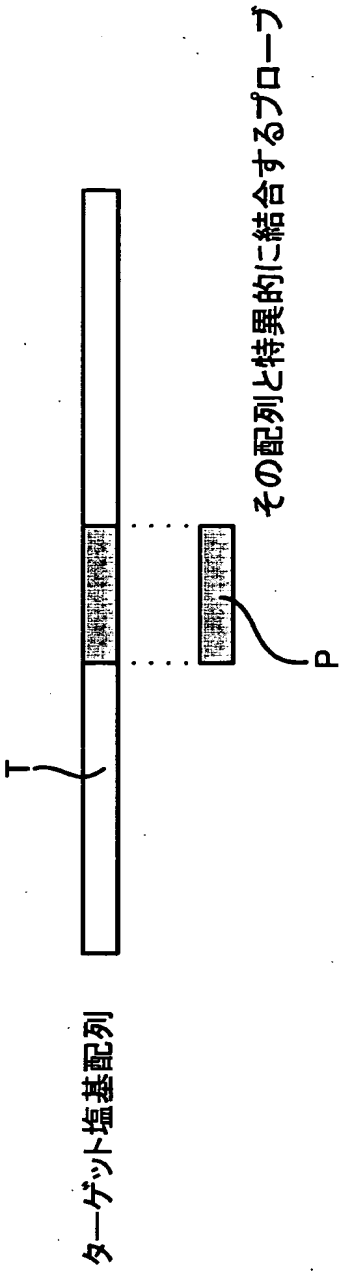
【図１６】

実施例				比較例 Smith-Waterman
相似度				
0.80	0.85	0.90	0.95	――――
405.47s	284.12s	136.20s	45.05s	1108.33s

【図１７】


比較例
(1108.33s)
計算時間(s)
450
400
350
300
250
200
150
100
50
0
70
75
80
85
90
95
100
相似度 (%)

【図１８】



【書類名】　要約書

【要約】

【課題】　塩基配列をスクリーニングするためのシステム、そのための方法、プログラム、記録媒体および塩基配列のスクリーニングを実行させるためのサーバを提供する。

【解決手段】　ターゲット塩基配列データおよびプローブ塩基配列とを格納するステップ（Ｓ１０）と、ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成し、ターゲット塩基配列とプローブ塩基配列との間に許容する最大の編集距離を格納するステップ（Ｓ１２）と、各塩基配列データおよび最大編集距離を各格納部から読み出してターゲット塩基配列データと相補配列データとを編集距離の大きな順に結合可能性を評価するステップ（Ｓ１４）と、結合性の評価の結果を記憶部に格納するステップ（Ｓ１６）とを含む。

【選択図】　図３

認定・付加情報

特許出願の番号	特願２００３－０５０９１６
受付番号	５０３００３１９５５３
書類名	特許願
担当官	伊藤　雅美　　２１３２
作成日	平成１５年　４月　９日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【住所又は居所】　神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権

【氏名又は名称】　上野　剛史

【復代理人】　申請人

【識別番号】　100110607

【住所又は居所】　神奈川県大和市中央林間３丁目４番４号　サクライビル４階　間山国際特許事務所

【氏名又は名称】　間山　進也

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　２００２年　６月　３日
　［変更理由］　　住所変更
　　住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
　　氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション